Geophysical Research Letters

COMMENT

10.1029/2022GL102530

Key Points:

- Scafetta (2022) contains errors in both of the statistical tests used that make the conclusions unsupportable

Correspondence to:

G. A. Schmidt,
gavin.a.schmidt@nasa.gov

Citation:

Schmidt, G. A., Jones, G. S., & Kennedy, J. J. (2023). Comment on "Advanced testing of low, medium, and high ECS CMIP6 GCM simulations versus ERA5-T2m" by N. Scafetta (2022). *Geophysical Research Letters*, *50*, e2022GL102530. https://doi.org/10.1029/2022GL102530

Received 14 DEC 2022
Accepted 28 JAN 2023

Author Contributions:

Conceptualization: Gareth S. Jones, John J. Kennedy
Writing – review & editing: Gareth S. Jones, John J. Kennedy

Comment on "Advanced Testing of Low, Medium, and High ECS CMIP6 GCM Simulations Versus ERA5-T2m" by N. Scafetta (2022)

Gavin A. Schmidt[1], Gareth S. Jones[2], and John J. Kennedy[3]

[1]NASA Goddard Institute for Space Studies, New York, NY, USA, [2]Met Office, Exeter, UK, [3]Exeter, UK

Abstract Scafetta (2022, https://doi.org/10.1029/2022gl097716) purports to test Coupled Model Intercomparison Project Phase 6 (CMIP6) climate models through a comparison of temperature changes over three decades. Unfortunately, the paper contains numerous conceptual and statistical errors that undermine all of the conclusions. First, no uncertainty is given for the observational temperature difference, making it impossible to assess compatibility with any model result. Second, the CMIP6 data are the ensemble means for each model, but the metric being tested is sensitive to the internal variability and so the full ensemble for each model must be used. When this is corrected, the conclusion that "all models with ECS > 3.0°C overestimate the observed global surface warming" is not sustained. Third, the statistical test in Section 2 would reject all models even in a perfect model setup given sufficient ensemble members, thus the second conclusion "that spatial *t*-statistics rejects the data-model agreement" is also not sustainable.

Plain Language Summary Comparisons of models and observations need to account from multiple sources of uncertainty in both the observations and due to the chaotic dynamics of the weather. The analyses in Scafetta (2022, https://doi.org/10.1029/2022gl097716) do not take either of these issues into account and thus the conclusions in that paper are not supportable.

1. Coupled Model Intercomparison Project (Phase 6) (CMIP6)

The Coupled Model Intercomparison Project Phase 6 (CMIP6) archive is a publicly available collation of climate model experiments performed by multiple groups around the world (Eyring et al., 2016). It includes simulations of the recent historical past, possible futures, and other idealized numerical experiments. The key characteristics of the archive is that there is (a) structural variability across models, (b) a quantification of initial condition uncertainty (since different initial conditions will give rise to different weather realizations), and (c) some exploration of forcing uncertainty and parametric uncertainty for some models. The historical hindcasts are run for 1850–2014, and are continued using the Shared Socioeconomic Pathways scenarios (2015–2100). For the purposes of this comment, we use the historical simulations paired with the SSP2-4.5 scenario.

We use CMIP6 simulations from the same source as Scafetta (2022), the Climate Explorer (ClimExp) from Koninklijk Nederlands Meteorologisch Instituut (Royal Netherlands Meteorological Institute) (KNMI). There are at present 175 individual simulations available from this portal that have historical and ssp245 continuations, from 36 models plus one physics variant which we treat as an independent model. One model listed in Scafetta (2022) does not have any ssp245 data available through ClimExp (note this is a limited subset of the full archive available through the Earth System Grid Federation). ClimExp offers multiple options for downloading the model data for instance, one simulation per model variant, or the ensemble mean per model variant—when more than one initial condition set is available, or a single simulation. In their Section 3, Scafetta (2022) claims to be examining single simulations from each model, however our replication shows that the results in their Table 1 are the ensemble means. We downloaded all the available simulations so that we could examine the impact of internal variability within each model. One model (FIO-ESM-2) does not have a documented climate sensitivity, so its 3 simulations are not used in our analysis (as with Scafetta (2022)), leaving 172 useable simulations. We take the Equilibrium Climate Sensitivity (ECS) values from Zelinka et al. (2020) (plus recent updates and corrections, (Hausfather et al., 2022)).

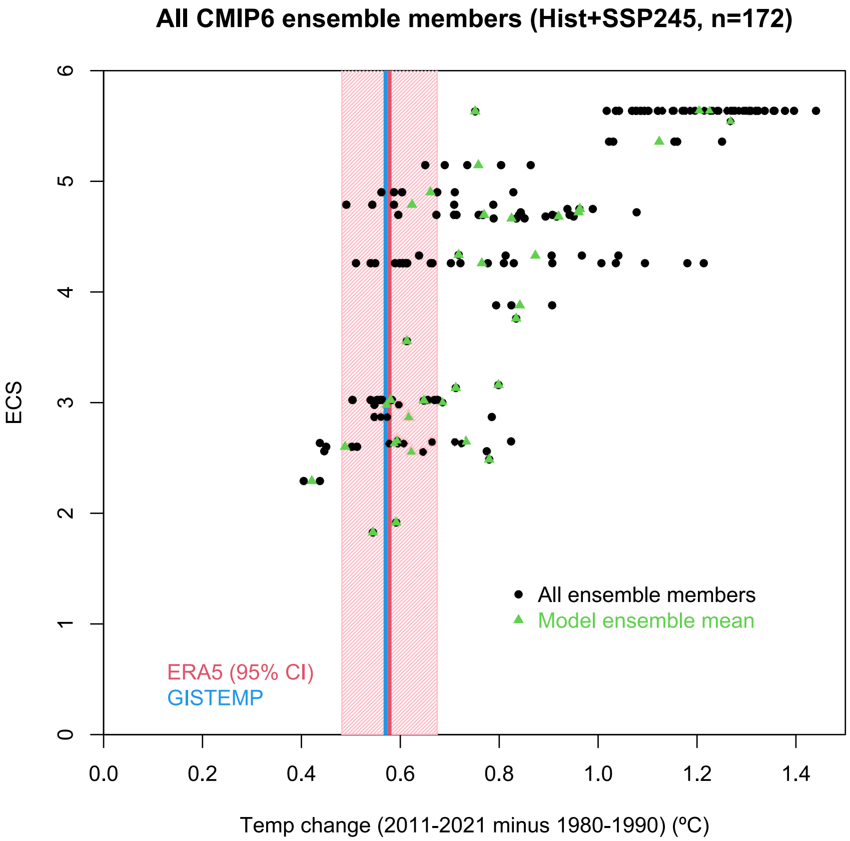

Figure 1. The difference between 1980–1990 and 2011–2021 in the global mean surface air temperature and the Coupled Model Intercomparison Project Phase 6 ensemble plotted against each model's Equilibrium Climate Sensitivity. Green triangles represent the model ensemble mean for each model or variant, while black dots represent (up to 25) other ensemble members. The pink shading represents the 95% uncertainty in the ERA5 estimate.

2. Global Mean Comparisons With Observations

We downloaded the global mean surface air temperature (SAT) from the European Centre for Medium Range Weather Forecasts Re-Analysis (version 5) (ERA5) (Hersbach et al., 2020; Simmons et al., 2021) directly from the Copernicus data store. We calculate the temperature difference between the two full 11-year periods 1980–1990 and 2011–2021. We note that this is not substantively different from the period used in Scafetta (2022) (January 2011–June 2021). We compare the same period in the models, again noting that this is not substantively different from the average of the 2011–2020 and 2011–2021 periods used in Scafetta (2022). These differences simplify the calculations without affecting the issues. Additionally, we compare the models to the GISTEMP observations (Lenssen et al., 2019) for the same periods.

Uncertainty in the expected temperature difference arises because of the random nature of internal variability (such as the timing of El Niño events), and the standard error can be estimated using the residuals of the annual data points that is,

$$\sigma_E = \frac{1}{\sqrt{N}}\sqrt{\sum\left(T_i - \overline{T}\right)^2}/\sqrt{N-1}$$

where T_i is the set of annual anomalies from 2011 to 2021 baselined to 1980–1990, and N is the number of years. We estimate that the mean and 95% confidence interval ($\pm 1.96 \times \sigma_E$) for the difference is then 0.58 ± 0.10°C for ERA5, and a very similar 0.57 ± 0.10°C for GISTEMP. The three decade period used in Scafetta's analysis is simply too short for internal variability to be ignored.

We summarize the results of the comparison in Figure 1, which can be contrasted with the right-hand panels in Scafetta's Figure 1.

First, even with just the ensemble means from each model, there are three models with ECS well above 3°C that can't be statistically distinguished from the observations. More importantly, looking at the full ensemble, we find that 49 ensemble members from 18 models are compatible with the ERA5 result. Of those 18 models, 9 of them have ECS above 3°C. This is in direct contradiction to the claims made in Scafetta (2022).

3. Spatial Comparisons and Statistical Test

Spatial patterns of change in models and observations are more affected by internal variability than the global mean. Thus even more care must be taken to compare like with like. It is a common error to compare the multi-model mean and its standard error with the observations. This test is essentially meaningless because we know a priori that they will not be equal (see Santer et al., 2008, for a discussion). Consider an ideal climate model, with perfect representation of the relevant physics and unlimited spatial and temporal resolution. An individual run of this model will not exactly match the observations because the initial conditions of the model run will not exactly match those of the real Earth. The model run will have the same forced response, but a different realization of the internal variability. Initial condition ensembles are therefore used to capture the statistical distribution of the effects of internal variability, with a better estimate of that distribution arising as more ensemble members were added. The standard error of the ensemble mean continuously decreases as more ensemble members are used, which means that the statistical test used in Scafetta (2022) is essentially guaranteed to reject a *perfect* model ensemble, and is therefore inappropriate.

Detecting a statistically significant difference between an individual model run held out from the ensemble and the mean of the remaining members would obviously not indicate a "model failure"; nor would it be an indication of an "inconsistency"—a model run cannot be inconsistent with the model from which it was generated. This provides a practical sanity check of any proposed statistical test; if a model ensemble is used to estimate the forced response, a test with 5% power should not reject individual held-out ensemble members more than 5% of the time on average. This was the key problem with test used in Douglass et al. (2008) that Santer et al. (2008) addressed. The test used by Scafetta (2022) also fails this check. In their Equation 2, the denominator contains a $\sqrt{N}$ term which means that as the number of ensemble members increases, so will the rejection rate. Thus the test is simply ill-formed.

A more appropriate test would be something that includes both observational uncertainty and ensemble spread, such as:

$$d = |\overline{T_m} - \overline{T_o}| / \sqrt{s\{< T_m >\}^2 + s\{T_o\}^2}$$

where $s\{<T_m>\}$ is the standard deviation of the model temperature differences T_m and $s\{T_o\}$ is the standard error of the observed temperature difference, respectively, following Santer et al. (2008) (their Equation 12 with a single model). This tests whether the observations are plausibly a sample from the distribution of model runs rather than for exact equality between the ensemble mean and the observations that physical considerations tell us will not be the case. In other words, it is a test of whether the observations are statistically exchangeable with the model runs. For the models with more than 2 ensemble members, 16 out of 24 models pass this test at the 5% confidence level, with sensitivities ranging from 2.3 to 5.1°C. It may also be important to account for the spatial correlation and the rate at which spurious results would be generated by chance with so many tests being performed at the gridbox level.

Given that an incorrect and misleading test (as has been long discussed in the literature) is being applied, we are confident that the conclusions drawn from the spatial tests in Scafetta (2022) are spurious or, at best, grossly exaggerated.

4. Additional Issues

There are a number of additional issues that, while minor relative to the two raised above, should nonetheless be acknowledged. First, it is important to note the forcing uncertainty over the historical period. For instance, the CESM2 model has been shown to have a noticeable sensitivity to changes in the source and frequency of biomass burning emission fields (Fasullo et al., 2022). A spurious global warming of up to 0.2°C was identified as a result of decadal mean biomass burning inputs being replaced by annually varying inputs, which led to a rectified effect on global temperature through a non-linear response to black carbon aerosols. Other forcings, such as ozone, or solar activity, are also imperfectly known, and this makes simple comparisons between the hindcasts and observations more complicated. Differences may arise between them not because of anything intrinsic to the model processes, but rather to the uncertainty in the drivers. Second, the number of ensemble members for many of the models is insufficient to estimate their forced response and magnitude of internal variability which limits the extent to which comparisons with those models will be informative.

In critiquing the tests in this particular paper, we are not suggesting that hindcast comparisons should not be performed, nor are we claiming that all models in the CMIP6 archive perform equally well. Indeed, there are multiple papers that demonstrate that CMIP6 models with high ECS values (above around 4.5°C) do not perform well in historical hindcasts (Ribes et al., 2021; Tokarska et al., 2020) or paleoclimate tests (Zhu et al., 2021). However, the claims in Scafetta (2022) are simply not supported by an appropriate analysis and should be withdrawn or amended.

Data Availability Statement

We use surface air temperature data from ERA5 (Hersbach et al., 2020) and GISTEMP (Lenssen et al., 2019). The CMIP6 ECS values are from Hausfather et al. (2022), and the annual global mean SAT from the CMIP6 models were downloaded from KNMI ClimExp (Trouet & Oldenborgh, 2013).

Acknowledgments

GSJ is supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. GAS is supported by the NASA Modeling Analysis and Prediction Program.

References

Douglass, D. H., Christy, J. R., Pearson, B. D., & Singer, S. F. (2008). A comparison of tropical temperature trends with model predictions. *International Journal of Climatology*, *28*(13), 1693–1701. https://doi.org/10.1002/joc.1651

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Fasullo, J. T., Lamarque, J.-F., Hannay, C., Rosenbloom, N., Tilmes, S., DeRepentigny, P., et al. (2022). Spurious late historical-era warming in CESM2 driven by prescribed biomass burning emissions. *Geophysical Research Letters*, *49*(2), e2021GL097420. https://doi.org/10.1029/2021GL097420

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., & Zelinka, M. (2022). Climate simulations: Recognize the 'hot model' problem. *Nature*, *605*(7908), 26–29. https://doi.org/10.1038/d41586-022-01192-2

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019). Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6307–6326. https://doi.org/10.1029/2018jd029522

Ribes, A., Qasmi, S., & Gillett, N. P. (2021). Making climate projections conditional on historical observations. *Science Advances*, *7*(4). https://doi.org/10.1126/sciadv.abc0671

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., et al. (2008). Consistency of modelled and observed temperature trends in the tropical troposphere. *International Journal of Climatology*, *28*(13), 1703–1722. https://doi.org/10.1002/joc.1756

Scafetta, N. (2022). Advanced testing of low, medium, and high ECS CMIP6 GCM simulations versus ERA5-T2m. *Geophysical Research Letters*, *49*(6), e2022GL097716. https://doi.org/10.1029/2022GL097716

Simmons, A., Hersbach, H., Muñoz Sabater, J., Nicolas, J., Vamborg, F., Berrisford, P., et al. (2021). Low frequency variability and trends in surface air temperature and humidity from ERA5 and other datasets. *European Centre for Medium-Range Weather Forecasts*. https://doi.org/10.21957/LY5VBTBFD

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, *6*(12). https://doi.org/10.1126/sciadv.aaz9549

Trouet, V., & Oldenborgh, G. J. V. (2013). KNMI Climate Explorer: A web-based research tool for high-resolution paleoclimatology. *Tree-Ring Research*, *69*(1), 3–13. https://doi.org/10.3959/1536-1098-69.1.3

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., et al. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019gl085782

Zhu, J., Otto-Bliesner, B. L., Brady, E. C., Poulsen, C. J., Tierney, J. E., Lofverstrom, M., & DiNezio, P. (2021). Assessment of equilibrium climate sensitivity of the Community Earth System Model version 2 through simulation of the Last Glacial Maximum. *Geophysical Research Letters*, *48*(3), e2020GL091220. https://doi.org/10.1029/2020gl091220